AFRL-RI-RS-TR-2009-217
Final Technical Report
September 2009

CONNOTATIVE MEANING OF MILITARY CHAT COMMUNICATIONS

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2009-217 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/s/	/s/
CHARLES G. MESSENGER, Chief	JOSEPH CAMERA, Chief
Information Understanding Branch	Information & Intelligence Exploitation Division
	Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* SEPTEMBER 2009	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* May 2008 – February 2009

4. TITLE AND SUBTITLE CONNOTATIVE MEANING OF MILITARY CHAT COMMUNICATIONS	**5a. CONTRACT NUMBER** In House (Mini-Grant)
	5b. GRANT NUMBER N/A
	5c. PROGRAM ELEMENT NUMBER 61102F
6. AUTHOR(S) Sharon M. Walter and Emily Budlong	**5d. PROJECT NUMBER** 230B
	5e. TASK NUMBER 8D
	5f. WORK UNIT NUMBER AH
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AFRL/RIED 525 Brooks Road Rome NY 13441-4505	**8. PERFORMING ORGANIZATION REPORT NUMBER** N/A
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFRL/RIED 525 Brooks Road Rome NY 13441-4505	**10. SPONSOR/MONITOR'S ACRONYM(S)** N/A
	11. SPONSORING/MONITORING AGENCY REPORT NUMBER AFRL-RI-RS-TR-2009-217

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA# 88ABW-2009-3614

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Automatic processing of military chat text in operational environments will be necessary to provide automated data collection, collation, and usage for tactical updates, post-mission operational analysis, and watch turnover. The informal nature of chat communications allows the relay of far more information than the technical content of messages. This AFRL in-house project combined components of the methodology applied in a Syracuse University project for IARPA's AQUAINT program with additional research activities to analyze databases of military chat. The project proposed to conduct a study of how humans recognize connotative cues expressing uncertainty, perception of personal threat, and urgency; formulate linguistic and non-linguistic means for recognizing such cues; develop algorithms to automatically perform that recognition, and evaluate the prototype recognition algorithms. The project built a matrix of speech "cues" representative of uncertainty, perception of personal threat, and urgency, but also applied maximum entropy analysis and a combined rule-based/statistical algorithm. Recall, precision and F-score measures for each methodology were determined.

15. SUBJECT TERMS
text chat, urgency, uncertainty, Maximum Entropy

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Sharon M. Walter
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	29	**19b. TELEPHONE NUMBER (*Include area code*)** N/A

Table of Contents

List of Figures

List of Tables

Acknowledgements

Our sincere thanks go to our collaborators at Syracuse University for their time, their guidance, and the access to TextTagger; and to the AFRL-Rome Research Site Library staff for their extraordinary assistance.

Also, thank you to AFRL/RI Chief Scientist Dr. John Bay
and the Technical Review Committee for the opportunity to pursue this research.

1.0 Introduction

Over the last five to seven years the use of chat in military contexts has expanded quite significantly, in some cases becoming a primary means of communicating time-sensitive data to decision makers and operators. During humanitarian operations with Joint Task Force-Katrina, chat was used extensively to plan, task, and coordinate pre-deployment and ongoing operations. The movement of amphibious craft for transporting personnel, equipment, and supplies ashore was coordinated and tracked through chat. During Operation Iraqi Freedom (OIF) the 4th Air Support Operations Group used chat continuously for Close Air Support (CAS) execution among four joint organizations. They provided Command and Control for all V Corps CAS missions and considered chat absolutely critical to mission accomplishment because it was the most expedient method of communication and allowed real-time collaboration [6]. In 2003 the U.S. Navy conducted a survey of chat usage by those on deployment for OIF. The majority of the 183 respondents indicated they used chat for over 7 hours per day, 6-7 days per week [12].

The use assessment of [Eovito; 2006] indicates that warfighters choose to use chat because it is fast, convenient, dependable, and efficient. The communication speed of chat is especially useful for tasking and re-tasking. Chat messages can be disseminated to everyone that would be involved in an operation and they can begin their preparation immediately. Collaboration among chat users doesn't require looking up electronic mail addresses, telephone numbers, or radio network identifications. Military chat users surveyed felt that without the use of chat, their situation awareness would be diminished, and information dissemination and coordination would be made more difficult. [Heacox, et all; 2004] noted that the use of chat instead of voice communications facilitates coalition operations since the problems derived from understanding accents and language fluency deficiencies are reduced. [Air Land Sea Application Center (draft); to be published: March 2009] points out that chat also provides a digital log of communications, allows operators to review missed posts, and allows more chat rooms to be monitored than voice channels. It further states, "IRC [Internet Relay Chat] enhances critical C2 capabilities through exponentially improved vertical and horizontal data communications by simultaneously transmitting and receiving C2 information to all participating and monitoring organizations across all echelons thus providing greater situational awareness resulting from increased information volume and reduced latency of information exchange." The document discusses the importance of chat as a command and control medium, not to replace existing formal communications but to enhance them by allowing timelier, accurate, and reliable planning, directing, and controlling of forces pursuant to the mission assigned.

Automatic processing of chat text has become necessary to provide for automated data collection, collation, and usage in new capabilities such as tactical updates, post-mission operational analysis, and watch turnover. The informal nature of chat communications allows the relay of far more information than the technical content of messages. Unlike formal documents such as newspapers, chat is often emotive. "Reading between the lines" to understand the connotative meaning of communication exchanges is now feasible and may become important for sounding alerts, for understanding behavior for after-action reviews, for participant identification verification, and for data collection and analysis.

2.0 Background

Most text analysis research to date has been on grammatical, well-formed text, such as articles in the Wall Street Journal. Analysis of chat text offers new challenges due to its dynamic nature. Chat messages often include misspellings, extra or missing capitalization, improper grammar constructs non-standard punctuation, abbreviations, interwoven conversations, and other unique characteristics. Some of the processing methodologies for linguistic analysis of grammatical text are being adapted for the special characteristics of text chat data (just three of many examples: [Srihari and Schwartzmyer; 2007], [Berube, et al; 2007] and a current project [Carpenter; 2008]). A number of other research studies are attempting to detect less concrete aspects of chat communications. Some of them have focused on detecting general emotion cues: ([Glazer; 2002], [Hancock, et al; 2007]). Other topics of chat study include the detection of empathy ([Pfeil and Zaphiris; 2007]), the detection of verbal irony ([Hancock, 2004]), and the detection of certainty (or confidence) and the measurement of the polarity of chat-detected sentiments, for example, negative/positive and favorable/unfavorable ([Liddy; 2004]).

The Center for Natural Language Processing (CNLP) of Syracuse University recently performed exploratory work for IARPA's Advanced Question and Answering for Intelligence (AQUAINT) project under a "blue sky" effort titled, "Understanding Connotative Meaning of Text." The project investigated how humans reach their understanding of the connotative meaning of text and developed some initial resources towards enabling computer understanding. Understanding the connotative meaning of text is currently beyond the capabilities of text processing systems but such understanding is necessary to enable more useful automatic intelligence exploitation.

This AFRL in-house project applied components of the research methodology applied in the Syracuse University AQUAINT project to two databases of military chat communications and expanded on its results. This project proposed to: (1) conduct a study of how humans recognize connotative cues expressing uncertainty, perception of personal threat, and urgency, (2) formulate linguistic and non-linguistic means for recognizing such cues, (3) develop algorithms to automatically perform that recognition, and (4) evaluate the prototype recognition algorithms.

3.0 Resources

3.1 Classified Chat Databases

Two chat databases from military exercises were available for this project. One database is from the 2006 Joint Expeditionary Force Experiment (JEFX-06); the other is a U.S. Navy chat database. (It may be that they are both from the same exercise, but we were unable to get clarification on that.) There are 38 chat rooms among the data, divided up by functional responsibility. Every entry is time-stamped.

In the future it’s expected that the number of standard chat rooms will be kept to a minimum and there will be a standardized naming convention for rooms within an area of operation. Rooms will be more formalized in creation, membership and communication content.

For the purposes of the technical activities within this project, three arbitrarily selected chat rooms were used as the test datasets. Dataset 1 had 204 lines of chat communication, Dataset 2 had 65 lines, and Dataset 3 had 190 line entries. After testing was completed, it appeared, from a glance through the data, that there were some minor differences in the “personality” of the rooms from which they were collected: Dataset 1 contained some casual conversation (more than the other two datasets), Dataset 2 was all technical exchange, and Dataset 3 was all technical and more intense than the other two.

3.2 Collaborator: Center for Natural Language Processing (CNLP) at Syracuse University

CNLP consultants on this project included the Dean of the School of Information Studies, Dr. Liz Liddy, Mike D’Eredita, Jaime Snyder, and Ozgur Yilmazel. Collaboration was through meetings in Syracuse, telephone communication and electronic mail.

4.0 Technical Activities

This section discusses the activities that were pursued for the project.

4.1 Contributory Activities

4.1.1 Phase II SBIR: "Extracting Time Critical Information from Dynamic Text"

A current Phase II Small Business Innovative Research Project with Stottler Henke (FA8750-07-C-0087) is developing software to extract domain-specific, time-critical information from text chat. Co-author, Ms. Budlong, attended a contract review on 20 August 2008 in Seattle, Washington to get supporting information for this in-house project. The presentation briefly discussed the potential use of "emotional indicators" in chat communications to determine truthfulness, but otherwise focused on applying current information extraction technologies as they have been applied to more formal texts and extending those technologies for recognizing dialog structure in series of chat communications. As with many projects focused on analyzing text chat, the unique factors of chat are mostly considered hindrances to their text extraction processes rather than important cues for deeper meaning.

4.1.2 Nellis CAOC

On 16 September 2008, co-author Ms. Budlong joined others in a visit to the Combined Air Operations Center (CAOC) at Nellis Air Force Base in Nevada, where CAOC personnel detailed their use of chat communications. The primary uses of chat at the Nellis CAOC are dynamic targeting and personnel recovery. For these applications, users are most interested in recognizing target identifiers, target locations, coordinates, and target location errors automatically. For most chat analysis applications it will be necessary to recognize the use of "possible," "probable," and "confirmed." All three were noted in the Navy chat databases as probabilities of events. The letter "c" was often used to confirm receipt of information. Guidelines under development ([ALSA; 2009]) will require that confirmations will specify the information that is being confirmed.

A common theme at operational sites is the lack of a common chat tool. Nellis personnel noted that InfoWorkSpace (IWS) and Mardam Internet Relay Chat (mIRC) each has its own pros and cons. As one example, IWS comes with more features but mIRC is cheaper.

CAOC personnel noted that keyword search is the current tool used to sort through chat data. It was noted that all information exchanged is not captured in chat: telephone, public announcement broadcasts, and person-to-person communications are also in use. "Whisper" chat, away from the formal chat rooms, frequently isn't logged also.

Chat conversations may be split across rooms. For example, a question may be asked in one room that requires the responder to seek out an answer. When the responder has an answer, possibly many minutes later, he may provide that answer to the questioner in a different room. Other issues include the need to "hunt down" people by checking in several rooms and inexperienced users utilizing the wrong rooms or passing information into incorrect chat rooms.

4.1.3 In-House Consultants

Upon recommendation from Mark Pronobis (Information Fusion and Understanding CTC Lead), we met with Robinson C. Ihle, an Intelligence Specialist with PAR Technology/Rome Research Corporation, on 8 September 2008. Rob works with the 152nd Air Operations Group in Syracuse, NY. The 152nd has bimonthly exercises during which some of the communication is by chat. There was further discussion at a later date with another Intelligence Specialist from PAR Technology: Anders Butler. Both gentlemen discussed current chat capabilities and functionality in military exercises. Exercises which include chat communications are performed almost monthly in Syracuse. There is a potential to visit during a CAOC exercise in the future, but personnel there have been too busy for such a visit during the course of this project.

4.1.4 Linguistic Inquiry and Word Count (LIWC)

A copy of the Linguistic Inquiry and Word Count (LIWC2007) software was purchased for $89.95 (http://www.liwc.net/liwcdescription.php) from laboratory funds. Development of LIWC was based on the notion that “the ways individuals talk and write provide windows into their emotional and cognitive worlds.” The LIWC website claims the software provides "an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals’ verbal and written speech samples."[1]

The software is designed to accept written or transcribed verbal text which has been stored as a digital file. With each text file, approximately 80 output variables are written as one line of data to a designated output file. This data record includes the file name, 4 general descriptor categories (total word count, words per sentence, percentage of words captured by the dictionary, and percentage of words longer than six letters), 22 standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.), 32 word categories tapping psychological constructs (e.g., affect, cognition, biological processes), 7 personal concern categories (e.g., work, home, leisure activities), 3 paralinguistic dimensions (assents, fillers, nonfluencies), and 12 punctuation categories (periods, commas, etc).

Code was implemented to calculate the percentage of words in each chat message that are found in the LIWC dictionary. When the code was run “right out of the box,” the messages and their LIWC word percentages were sorted by percentage and output to an Excel file. Interestingly, casual conversation text tended to have high percentages of LIWC dictionary words (at least 70%) and the on-topic technical chat messages had lower percentages (around 30%). In a manner of speaking, LIWC separated “the chaff from the wheat.” This simple result might be important for some automatic chat processing but guidelines being developed for military chat communications [1] dictate that chat room owners will prevent off-topic and inconsequential chat from occurring in their rooms.

[1] Interestingly, the webpage http://wordwatchers.wordpress.com/ shows the results from applying LIWC to analyze candidate speeches in the 2008 Presidential Election.

Words that make up the LIWC word categories of “anxiety,” “tentative” and “certainty” were reviewed as potential cues for urgency, uncertainty and perception of personal threat, however the breadth of words in each category and the brevity of each message severely dilute the applicability of LIWC for this project. It appears that LIWC may provide some insight about the general tone of a larger conversation, but fails to provide significant information across short communications. Our manual review of the chat data did indicate that there are words and phrases that seem to connote our focus stresses. They are included in our Results (Section 5.0).

4.2 Rule-Based Analysis: Identifying and Detecting Cues

Our research proposal was focused on recognizing connotations of uncertainty, perception of personal threat, and urgency in military chat data. We proposed to look for cues within chat messages and within the context established by surrounding chat communications. The following content of this section discusses the manual review of the data and cues that were potential sources of information.

4.2.1 Manual Review of Chat Data for Cues of Connotative Meaning

A manual review of the data was performed first, to attempt to understand what cues were indicative of our focus stress types. The data was reviewed for linguistic and non-linguistic cues. An example of a linguistic cue would be the use of particular words and phrases. Non-linguistic cues under investigation included: terse/lengthy responses (presuming that lengthy responses are very rarely used under circumstances of urgency or uncertainty), the use of capitalization, punctuation (including ellipsis), abbreviations, irregular spelling and metadata values.

One sign of emotion in general chat communications is the use of capitalized words as a means of indicating high emotions or angry screaming. In our own laboratory experience, we found that the use of capitalization in military chat is used for catching attention or alerting other chat participants to important information; it is rarely, if ever, used for “screaming.” Punctuation has been referred to as the ‘prosody of online communication’ [10], providing the equivalent of speech intonation in text to relay connotative meaning. In many ways chat communications are similar to transcribed spoken dialogue. For instance, they often contain interjections, such as “ah!” and “drat!” In a distinction from general chat, however, based on our experience, military chat interjections rarely include the identifying punctuation.

Abbreviations that are common in general chat communications, such as “msg” (“message”) and “thx” (“thanks”), were present in our datasets along with an additional set of chat abbreviations that are specific to military communications (for instance, “w/u” to mean “wheels up”). [ALSA; 2009], a developing document to facilitate coordination of military chat use, recommends avoiding “civilian convenience” abbreviations and includes a table of standardized chat terminology. Some abbreviations are easily recognizable and commonly used in civilian chat, for example “arr” (arrived), “neg” (negative), “unk” (unknown); other abbreviations are unique to the military domain. The databases used for this project were not limited by the restrictions by [ALSA; 2009].

Irregular spelling could be accidental misspelling, potentially due to rushed typing, or a purposeful expression (ex. “riiiight” as an indication of the mental dawning of agreement, as opposed to “right” as an indication of simple, immediate agreement). As in transcribed speech, ellipsis, the trail of dots that indicates an incomplete thought or an omission of words (ex. “Well, if that’s so…”), is very common in both general and military chat communications. Metadata values, considered as sources for cues, could be the identity of chat participants and the time between exchanges. The identification of the chat participant and the associated temporal label at the beginning of each message are distinctive characteristics of chat communication that are not available in formal texts and can offer valuable information for processing systems. For example, for our purposes in this project, knowledge of the functional role and status of particular speakers could have been important input to the determination of connotative intent. We did not, however, have that information. As it turned out, the temporal component of the metadata was of no use in recognizing either urgency or uncertainty for this project. Exchanges were sometimes made across more than one room (question is made in one room, answer is given in another), communication seemed lax with lengthy response times (possibly due to the fact that it is data from an exercise), and dialogue sequences were difficult to untangle.

Our chat databases included a couple of emoticons (smileys) in a few casual conversation inputs, but it is unlikely that smileys will be used in formal military chat under intense field conditions.

4.2.2 Cues for Uncertainty and Urgency

We found uncertainty and urgency cues to be quite subtle, and found no exemplars of “perception of personal threat” within the data. Possible explanations for the latter are that: (1) evidence of personal threat was too subtle for our detection efforts, (2) exemplars of perception of personal threat don’t exist in our datasets because the data was recorded during an exercise rather than a real battlefield event, or (3) perception of personal threat would not be present in the military chat domain whether in an exercise or a real event because chat participants don’t feel any personal threat. Since the “perception of personal threat” could not be recognized within any database entries, uncertainty and urgency became the foci of the study.

It appeared that cues for uncertainty and urgency gave a varying level of confidence of the existence of uncertainty/urgency, so confidence scores from 1 to 5 (5 being an indication of the highest confidence) were attached to each cue. Table 1 lists the uncertainty and urgency cues and scores that were developed. For each cue, an explanation of the cue (sometimes with an example or two), the connoted meaning(s) (uncertainty and/or urgency), a confidence score and the proposed software implementation for automatically detecting the cue, are listed. Note that examples are very simple and intended only for illustration of the syntax being described.

In our definition of uncertainty, we were looking for messages expressing more than a simple need for information. For example, the single message "What time are we striking?" with no other questions near it would be considered a simple request for information. However, when there are more questions in the same message or in consecutive messages, the person(s) involved is (are) more likely to be demonstrating a state of confusion (that is, uncertainty).

Urgency seemed to be fairly cut and dry, and dependent on keywords. Messages that end with "ASAP", "immediately", or "press" were very likely to be expressing urgency. Messages ending with "now" are a little more difficult, as the message could be "Get this done now", or it could be "I'm working this now". The first may be expressing urgency; the second is more of a status update. Other than these keywords, we did not find any syntax that seemed to express urgency. As noted earlier, capitalization did not provide significant evidence of urgency in our data, as it is used largely just to catch the attention of the intended recipients. The use of capitalized "NO" to indicate urgency was a rare exception. Exclamation points were rarely used, and usually did not convey urgency.

Confidence scores are added when more than one cue is found in relation to a chat message. Thus, Example 1 would achieve a total confidence score of 10 for uncertainty for having two questions in one message (5 points), multiple question marks at the end of a question (3 points), and questions and ellipsis in the message (2 points).

Example 1:
Person A: "Are we striking at 1400? Where are we striking?? I can't access the info…"

Example 2:
Person A: "Are we striking at 10??"
Person B: "I don't know…do you know where we're striking?"

Example 2 would get 4 points for consecutive questions across speakers, 3 points for multiple question marks, and 2 points for question and ellipsis in a single message, for a total uncertainty score of 9.

4.2.3 TextTagger

Collaborator Syracuse University Center for Natural Language Processing provided a copy of TextTagger, their rule-based information extraction system that analyzes unstructured text for lexical, syntactic and semantic information. TextTagger breaks input text files into sentences, brackets meaningful phrases such as temporal concepts and named entities, assigns part-of-speech tags to words and phrases, and tags entities, events, and relations. Many of the cues that we determined to be relevant to indentifying connotative meanings are not tagged in TextTagger. Currently there are no annotation guidelines widely accepted by the research community for the unique aspects of text chat. An attempt to create annotations of chat phenomena is documented in [Creswell; 2006], but the work was never considered polished enough for broad distribution.

4.2.4 Software Program for Cue Analysis

It was determined that some of the cues could be captured by regular expressions (recognizable patterns that can be interpreted into software code) in TextTagger, some would require dialogue processing because the cues were found in multi-sentence sequences, and some would be best processed at the conclusion of TextTagger processing (see Table 1). Rather than pursuing all three modes of analysis, including the difficult work of adding processing implications to TextTagger, and due to the time-constraints of this project, a separate Java software program was developed to perform the recognition of cues. One of the cues (#7: “Which [noun]?”) would require recognition of the classification of a word as a ‘noun’ by a software parser and, although the rule is probably pertinent, it would not have been applied very many times in relation to the time and effort it would have taken to implement it. Therefore, Rule #7 is not implemented in the code.

The results of applying this cue recognition program are provided in Section 5.0.

4.3 Statistical Analysis: Maximum Entropy (“MaxEnt”)

During the course of this project, maximum entropy was suggested by co-workers to be a potential statistical analysis mode of analysis that could be applied to our data. Maximum entropy is a statistical modeling technique in which a dataset from a seemingly random process is used to make predictions about future data output. For this project, OpenNLP group's Maximum Entropy package [19], open source code written in Java, was given a set of data which was tagged with indications of our conclusions about their connotative content as a training set. Training data was derived from chat rooms other than the three testing datasets, from the same data source and event. Approximately 20 samples representing each of uncertainty, urgency and other were used for training. When the trained system was then applied to the test data, it automatically classified chat statements as containing cues of urgency, of uncertainty, or other.

We were interested in determining the effect of stop words[2] in the datasets on MaxEnt analysis. The stop word list used is from: http://www.lextek.com/manuals/onix/stopwords1.html [18]. Datasets containing stop words and datasets with stop words removed were developed, and MaxEnt was run on both datasets. Interestingly, datasets without the stop words had a much lower classification accuracy than datasets with the stop words. Results included in Section 5.0 are from analysis when stop words were not removed.

[2] Stop words are the very common words, such as “a,” “an,” and “the,” that are often eliminated from text resources before information retrieval operations are performed.

Table 1: Cues developed by manual review of data

Cue Description		Explanation	Connoted Meaning	Points (1 - 5)	Proposed Software Implementation
1	Two or more questions in one message.	One speaker, one message, with two or more questions. More questions within one message indicate more uncertainty.	uncertainty	5	Use TextTagger to identify sentence boundaries and check for question mark at the end; count question marks. Overkill to use TextTagger.
2	Questions with an option.	A question that gives a choice. Example: “Should target A be our priority, or is target B more important?”	uncertainty	4	Very difficult for automatic processing. Not available with TextTagger.
3	One speaker with two or more questions in consecutive messages.	Example: Person A: “Are we striking at 1400?” Person B: “affirmative, strike at 1400.” Person A: “copy, what are the cords for the strike?” Person B: “56N 138W”	uncertainty	4	Same answer as #1: TextTagger and analyze TextTagger ouput.
4	Two or more consecutive questions across speakers.	In consecutive messages, regardless of speaker, each message has at least one question. Example: Person A: “Are we striking at 1400?” Person B: “Is the location still 56N 138W?”	uncertainty	4	Same answer as #1: TextTagger and analyze TextTagger ouput.
5	Multiple question marks at the end of a question.	More question marks usually mean more uncertainty.	uncertainty	3	Same answer as #1: TextTagger and analyze TextTagger ouput.

6	“understand” and a question mark in a message.	Example: “I don’t understand. Weren’t we targeting A?”	uncertainty	3	Could be done with regular expression as TextTagger post-process.
7	“Which [noun]?”	Self-explanatory.	uncertainty	3	TextTagger annotates parts-of-speech, including nouns. Post-process by looking for “which” followed by a noun.
8	Question and ellipsis in one message.	Examples: “What time are we striking? I lost the info…” “Do you know who we are looking for…?”	uncertainty	2	Could be done with regular expression as TextTagger post-process.
9	Ellipsis	Sentence within a message ends with “…”	uncertainty	1	TextTagger detects sentences. Post-process by looking for sentence ending with ellipsis.
10	“ASAP,” “immediately,” or “press” at the end of a sentence.	Self-explanatory.	urgency	4	Could be done with regular expression as TextTagger post-process.
11	“now” at the end of a sentence.	Self-explanatory.	urgency	3	TextTagger detects sentences. Post-process by looking for sentence ending with “now.”
12	Capitalized NO.	Example: “NO impact”	urgency	3	Could be done with regular expression as TextTagger post-process.
13	“hot” somewhere in the message.	Example: “Going hot with target A”	urgency	2	Could be done with regular expression as TextTagger post-process.

4.4 Combined Rule-Based and Statistical: Parallel Analysis

In a final data analysis, software code was written to combine MaxEnt and our Cue Table for a parallel analysis. Cue Table confidence scores were divided by 5 in order to force a basis for comparison with MaxEnt. Granted, this was a very gratuitous translation, but the final results were surprising. For each message within each of three datasets, the decisions of MaxEnt and our Cue Table are considered together and final results are produced as shown in the table below. So, for example, if the decisions of MaxEnt and the Cue Table are the same, then the final decision of the combined algorithm is that same decision. If MaxEnt indicates Urgency and the Cue Table indicates Uncertainty, then the final decision will be determined by the highest confidence score between them.

Table 2: Parallel Analysis with MaxEnt and Cue Table

		Cue Table/5		
		Urgency	Uncertainty	Other
MaxEnt	Urgency	Urgency	Highest Confidence Scorer	If MaxEnt Confidence Score > .6, Urgency; Otherwise, Other
	Uncertainty	Highest Confidence Scorer	Uncertainty	If MaxEnt Confidence Score > .6, Uncertainty; Otherwise, Other
	Other	Highest Confidence Scorer	Highest Confidence Scorer	Other

Note that for MaxEnt Urgency or Uncertainty, with Cue Table decision of Other, the final decision is based on the MaxEnt confidence score. If the MaxEnt confidence score is greater than .6, then the final decision will match the MaxEnt decision for that message. If the MaxEnt confidence score is less than or equal to .6, then the final decision will be Other.

5.0 Results

5.1 Information Extraction Metrics

Recall and precision are commonly used measures applied to tasks similar to this project. The meanings of recall and precision can be clarified by the Venn diagram of Figure 1 in which the circle on the left represents all of the information of interest in the dataset (that is, the ground truth) and the circle on the right represents the information selected by the software analysis. Therefore, the intersection, a, represents the information of interest that was correctly identified by the automatic analysis. The rectangle represents the entire dataset (the Universe).

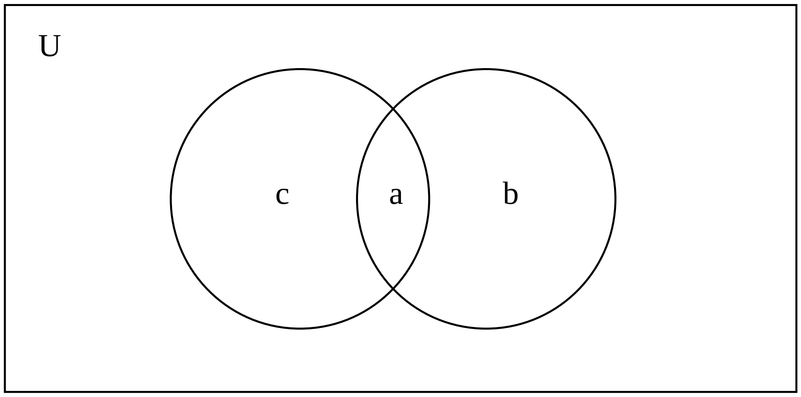

Figure 1: Venn Diagram. The circle on the left (a + c) represents all of the information of interest. The circle on the right (a + b) represents the information selected by an automatic analysis.

A recall measure represents the amount of correct, relevant information that was identified in comparison to the total amount of relevant information (that is, the ground truth) within the dataset. The equation for recall, as represented by the diagram of Figure 1, is:

$$\text{Recall} = \frac{a}{a+c} \qquad (1)$$

A recall score of 1.0 would mean that all of the relevant information was correctly identified. It is a measure of the completeness of the data identified. Note that selecting all of the information in a dataset would measure out as perfect recall, but the result would be of no use.

A precision measure represents the amount of information that was correctly identified in comparison to the amount of all of the information that was identified by the analysis. The equation for precision is shown, below. A perfect precision score of 1.0 means that all of the information selected as being relevant, is actually relevant. Note that this wouldn’t necessarily mean that all of the relevant information in the dataset has been detected.

$$\text{Precision} = \frac{a}{a+b} \qquad (2)$$

The F-measure is the weighted harmonic mean of precision and recall, useful for comparing capabilities of systems as a single measure. For some analysis applications, one of either recall or precision may be more highly valued and that would determine the weight of each of them in the calculation of the F-measure. The research metric traditionally used is the balanced F-score, with evenly weighted recall and precision (often indicated as F1):

$$\mathrm{F} = \frac{2\ x\ (precision\ x\ recall)}{precision + recall} \qquad (3)$$

Table 3 (page 14) shows the values for precision, recall and the balanced F-score for analysis of the three datasets and for their combination as a single set of data for each of the analysis methodologies of: cue analysis, maximum entropy and parallel analysis. Scores are multiplied by 100, as had been the practice of DARPA funded Message Understanding Conference (MUC) evaluations.

5.2 Analysis of Results

Our attempts to recognize urgency failed, but have led to numerous potential directions for continued research. The cues we thought we observed were vague to begin with (see Section 4.2.2), focusing on keyword matching. After-test review of the data showed that some overgeneration[3] was caused by rule 12 of Table 1 that looked for a capitalized NO as an indication of urgency because the rule matched numerous references to a chat participant whose function name included the word NO within it. Improving the recognition of urgency would require a completely new look at the problem. Perhaps urgency would be better recognized if the time between chat entries and a count of misspelled words could be used as cues.

The results for recognizing uncertainty were also disappointing but the scores and further scores achieved by manipulation of data rules seem to point towards some validation of the project's direction. It may be that uncertainty was more readily detectable for this application due to the inclusion of novice participants in the recorded activities.

The cue analysis recall score for all of the data as a single dataset was 40.48. The precision score was 46.58. Manual review of the labeling indicated that a large amount of the overgeneration by the cue analysis algorithm was due to one particular rule – the ellipsis rule (rule #9 in Table 1). The rule labeled every chat entry containing ellipsis to be representative of uncertainty. Rule 8, marking an entry as uncertain if it contains a question *and* an ellipsis, was correct a larger percentage of the time. Eliminating rule 9 increased the precision significantly (to 75.00; as shown by parenthesized entry in Table 3). Eliminating that rule, of course, reduced the recall value because some of the recognitions would have been valid, but the reduction in recall was less significant than the increase in precision. It may be that further investigation could refine a rule or ruleset for recognizing uncertainty in chat messages with ellipsis.

[3] Overgeneration is a somewhat out-of-vogue measure. It looks at extraction results from a negative perspective rather than from the positive perspective of recall and precision.

Rule #7 (“Which <noun>?”) was the only rule developed that would require parsing or part-of-speech analysis. With further investigation, or within other chat databases, deeper grammatical analysis might produce more and/or stronger cues of uncertainty. As noted in Section 4.2.4, this rule was not included in the evaluations.

It was observed during the manual data review that phrasing of messages indicated uncertainty. For example, the message “Do you know if we should be tracking target A?” conveys more uncertainty than the message “Where is target A?” Although both are requests for information, the first has a tentative tone to it whereas the second has a more business-like tone. Consideration of phrasing of messages was left to be considered in future work.

Table 3 shows that maximum entropy analysis recall of uncertainty was significantly higher than that of our cue analysis, demonstrating that MaxEnt was able to recognize many more of the chat entries presenting uncertainty. MaxEnt’s precision, however, was, in most cases, lower than that of cue analysis. Statistical analyses like MaxEnt can often be improved (to a point) with additional training. It would be interesting to determine the amount of training data that would have returned the best possible results.

Table 3: Uncertainty
(Parenthesized entries are results of cue analysis without Rule #9.)

	Cue Table			MaxEnt			Parallel		
	Recall x 100	Precision x 100	F-score x 100	Recall x 100	Precision x 100	F-score x 100	Recall x 100	Precision x 100	F-score x 100
Dataset 1	39.47 (23.68)	35.71 (75.00)	37.50	78.95	45.45	57.69	57.89	59.46	58.66
Dataset 2	25.00 (25.00)	50.00 (50.00)	33.33	100.00	32.00	48.48	75.00	54.55	63.16
Dataset 3	44.74 (42.12)	62.96 (80.00)	52.31	60.53	35.94	45.10	50.00	70.37	58.46
All Data	40.48 (32.14)	46.58 (75.00)	43.32	72.72	39.35	51.07	55.95	62.67	59.12

Table 4: Urgency

	Cue Table			MaxEnt			Parallel		
	Recall x 100	Precision x 100	F-score x 100	Recall x 100	Precision x 100	F-score x 100	Recall x 100	Precision x 100	F-score x 100
Dataset 1	16.67	20.00	18.18	16.67	6.25	9.09	16.67	14.29	15.39
Dataset 2	0.00	0.00	0.00	40.00	25.00	30.77	0.00	0.00	0.00
Dataset 3	100.00	6.90	12.91	50.00	3.85	7.15	100.00	0.00	2.00
All Data	23.77	8.82	12.87	30.77	8.00	12.70	23.08	12.50	16.22

It should be noted that our application of MaxEnt was not able to detect connotative meanings for which evidence is provided within a dialogue, that is, across multiple chatlines, as identified in our cue analysis rules #3 and #4.

Parallel analysis, described in Section 4.4, was implemented upon realizing that MaxEnt recall scores were much better than cue analysis, and cue analysis precision scores were a bit better than MaxEnt. This algorithm achieved an overall recall score between that of MaxEnt and the cue table (55.95), and its precision and F1 scores were higher than those of MaxEnt and the cue table. It may be that using a different threshold score for the confidence level of MaxEnt in the parallel algorithm would result in better performance. We ran the parallel algorithm with thresholds of .5 and .6 and found that .6 yielded better performance. Further work would have to be done to determine if that is the optimal threshold value.

Although we were initially disappointed by the results for the three analysis methods, it should be noted that the scores for uncertainty recognition were comparable to scores achieved in very early analyses such as the third Message Understanding Conference of 1991. Figure 2 shows the relationship of the cue table, MaxEnt, and Parallel results among the results of the dry run of the named entity task of the first Message Understanding Conference that had a dedicated named entity portion (MUC-6).

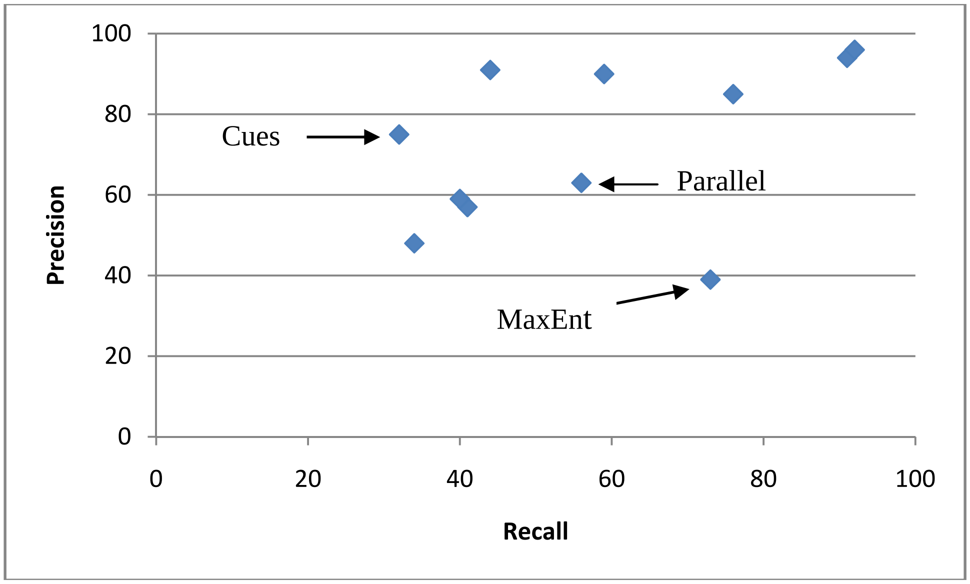

Figure 2: Cue, MaxEnt and Parallel scores among MUC-6 Named Entity scores.

Recognizing urgency and uncertainty is much more subjective than recognizing named entities. Urgency and uncertainty can be expressed in varying degrees throughout a conversation and cues for them are not as definitive as for named entities. Perhaps a sliding scale with spans of varying length representing a general consensus of the degree of urgency or uncertainty, rather than single-point measures, would be more appropriate for measuring such subjective concepts. The degree of expression of such connotative meanings could be monitored across time over the course of a conversation.

Context did appear to be useful for determining whether chat messages expressed urgency or uncertainty. Dialog analysis and deeper grammatical analysis would be useful for finding less explicit meanings of chat messages more accurately than the rule-based and simple statistical approaches used for this project.

6.0 Conclusions

Automatic recognition of connotative meaning could help military chat room administrators determine when team members are overwhelmed, recognize when a confused team member needs an information clarification, and provide an alert when something needs their immediate attention. However, the tactics, techniques, and procedures for military use of chat are still developing. For example, the guidelines in [ALSA; 2009] suggest that chat room administrators prevent off-topic and inconsequential chat from occurring. As a result, some of the conclusions of this project are more applicable to chat communications in general internet communities than in military communications. As military chat becomes less free-form the cues for concepts such as urgency and uncertainty may become even more subtle.

This document describes potential directions for future research for the detection of connotative meaning, including counting misspelled words and monitoring the time between chat entries as a means of recognizing urgency, and improving the ellipsis rule in order to recognize uncertainty. Further research for uncertainty, urgency and other connotative meanings could include more thorough dialogue analysis, deeper grammatical analysis, additional statistical training and determination of optimal threshold values for statistical analysis applications, and investigation of knowledge structures for representing degree and consensus of connotative meaning.

The eventual automatic processing of chat-formatted text and all of the subtleties it presents requires the research community to develop annotations for the peculiarities of text chat and annotated chat databases. [Creswell, et al; 2006] is a guide for marking up chat text data for computer analysis using an annotation tool called GATE. The guidelines in the document are intended for documentation of idiosyncrasies of chat text to enable better performance by information extraction software systems. Annotated phenomena are divided into low-level and high-level categories. High-level categories describe discourse-level properties that, for the most part, were not relevant for the in-house project documented in this paper. Low-level phenomena described included typographic errors, non-standard orthography, non-standard punctuation, and ungrammatical constructions. Many of these low-level phenomena are indicators of the connotations that this project sought to detect. Although the document provides some inroads to the research required for automatic chat text analysis, it has not been circulated into the wider research community. [Forsyth; 2007] is among other recent projects to develop chat annotations and annotated chat databases. Their annotation notation starts with the Penn Treebank part-of-speech tagging, then adds notations for dialog acts, misspellings and chat-unique features such as ellipsis.

In working this project we learned a great deal about military chat communication and operational chat utilization, on-going research in the analysis of dynamic text, and current text analysis tools and their applicability to dynamic texts.

7.0 References

[1] Air Land Sea Application (ALSA) Center, “IRC: Multi-Service Tactics, Techniques, and Procedures for Internet Relay Chat for Command and Control Operations (Coordinating Draft),” March 2009.

[2] Berube, Christopher D., Janet M. Hitzeman, Roderick Holland, Robert L. Anapol and Stephen R. Moore, “Supporting Chat Exploitation in DoD Enterprises,” Proceedings of the 12th International Command and Control Research and Technology Symposium (CCRTS), Newport, RI, 2007.

[3] Carpenter, Tamitha, “Extracting Time Critical Information from Dynamic Text,” Briefing Charts from 20 August SBIR Phase II Status Review Meeting, Unpublished, 2008.

[4] Creswell, Cassandre, Nicholas Schwartzmeyer, and Shane Axtell, “Dynamic Text Sources (chat) Annotation Manual,” August 2006, unpublished.

[5] Creswell, Cassandre, Nicholas Schwartzmeyer and Rohini K. Srihari. "Information Extraction for Multi-Participant, Task-Oriented, Synchronous, Computer-Mediated Communication: a Corpus Study of Chat Data", in *Proc. IJCAI-2007 Workshop on Analytics for Noisy and Unstructured Text Data*, Hyderabad, India, January 2007, Pg. 131-138.

[6] Eovito, Bryan A., “An Assessment of Joint Chat Requirements from Current Usage Patterns,” Naval Postgraduate School, Monterey, CA, DTIC: ADA451327, June 2006. http://handle.dtic.mil/100.2/ADA451327

[7] Forsyth, Eric Nielsen, “Improving Automated Lexical and Discourse Analysis of Online Chat Dialog,” September 2007.

[8] Gajadhar, Joan and John Green, “An Analysis of Nonverbal Communication in an Online Chat Group,” Open Polytechnic of New Zealand Working Papers, 2003. http://repository.openpolytechnic.ac.nz/eserv.php?pid=openpoly:44&dsID=res_wp203gajadharj1.pdf (Accessed 26 September 2008)

[9] Glazer, Courtney. “Playing Nice with Others: The Communication of Emotion in an Online Classroom,” Presented at 9th Annual Distance Education Conference, Austin, TX, 2002. http://www.scholarlypursuits.com/dec_comm.pdf.

[10] Hancock, J.T., “Verbal Irony Use in Computer-Mediated and Face-to-Face Conversations,” Journal of Language and Social Psychology, 23, 2004. Pg. 447-463.

[11] Hancock, J.T., Christopher Landrigan and Courtney Silver, “Expressing Emotion in Text-Based Communication,” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose, CA, 2007, Pg. 929-932. http://delivery.acm.org/10.1145/1250000/1240764/p929-hancock.pdf?key1=1240764&key2=6694994321&coll=GUIDE&dl=GUIDE&CFID=22634219&CFTOKEN=26358749

[12] Heacox, Nancy J., Ronald A. Moore, Jeffrey G. Morrison and Rey F. Yturralde, “Real-Time Online Communications: ‘Chat’ Use in Navy Operations,” 2004 Command and Control Research and Technology Symposium, June 2004. http://stinet.dtic.mil/cgi-bin/GetTRDoc?AD=ADA465828&Location=U2&doc=GetTRDoc.pdf

[13] Liddy, Elizabeth D., “Extraction of Elusive Information from Text,” Proceedings of the International Association of Science and Technology for Development Conference on Knowledge Sharing and Collaborative Engineering, St. Thomas, U.S. Virgin Islands, 2004.

[14] Liddy, Liz, Ozgur Yilmazel and Mike D'Eredita, "Understanding the Connotative Meaning of Text: A Blue Sky Project," AQUAINT-3 Kick Off Meeting, 31 October – 2 November 2006.
http://cnlp.syr.edu/presentations/slides/liddy.aquaint3.kickoff.pdf

[15] Pfeil, Ulrike and Zaphiris, Panayiotis, Patterns of Empathy in Online Communication, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (San Jose, California, USA, April 28 - May 03, 2007). CHI '07. ACM Press, New York, NY, Pg. 919-928, 2007.

[16] Srihari, Rohini K. and Nicholas Schwartzmyer, "Adapting Information Extraction Technology to Computer-Mediated Dynamic Text Data," AFRL-IF-RS-TR-2007-94.

[17] "Linguistic Inquiry and Word Count," http://www.liwc.net/. Accessed 28 August 2008.

[18] "Onix Text Retrieval Toolkit: API Reference," http://www.lextek.com/manuals/onix/stopwords1.html. Accessed 21 July 2008.

[19] "OpenNLP MAXENT," http://maxent.sourceforge.net/. Accessed 28 August 2008.

[20] "Penn Treebank II Tags," http://bulba.sdsu.edu/jeanette/thesis/PennTags.html. Accessed 29 October 2008.

[21] "Summary Score Reports for MUC-6 Dry Run of the Named Entity Task," http://www.cs.nyu.edu/cs/faculty/grishman/score-summary.NE.11may95. Accessed 23 January 2009.

8.0 List of Acronyms

AFRL	Air Force Research Laboratory
ALSAC	Air Land Sea Application Center
AQUAINT	Advanced Question and Answering for Intelligence
ASAP	As soon as possible
C2	Command and Control
CAOC	Combined Air Operations Center
CAS	Close Air Support
CNLP	Center for Natural Language Processing
CTC	Core Technical Competencies
DARPA	Defense Advanced Research Projects Agency
IARPA	Intelligence Advanced Research Projects Activity
IRC	Internet Relay Chat
IWS	InfoWorkSpace (a chat tool)
JEFX	Joint Expeditionary Force Experiment
LIWC	Linguistic Inquiry and Word Count (commercial software)
MaxEnt	Maximum Entropy
mIRC	Mardam Internet Relay Chat
MUC	Message Understanding Conference
OIF	Operation Iraqi Freedom
SBIR	Small Business Innovative Research